Huib Kouwenhoven*, Mirjam Ernestus and Margot van Mulken

Register variation by Spanish users of English: The Nijmegen Corpus of Spanish English

Abstract: English serves as a *lingua franca* in situations with varying degrees of formality. How formality affects non-native speech has rarely been studied. We investigated register variation by Spanish users of English by comparing formal and informal speech from the Nijmegen Corpus of Spanish English that we created. This corpus comprises speech from 34 Spanish speakers of English in interaction with Dutch confederates in two speech situations. Formality affected the amount of laughter and overlapping speech and the number of Spanish words. Moreover, formal speech had a more informational character than informal speech. We discuss how our findings relate to register variation in Spanish.

Keywords: ELF, non-native, register variation, formality, English, Spanish

DOI 10.1515/cllt-2013-0054

1 Introduction

English is the most widely used means of communication during international encounters (e.g. De Swaan 2001). The study of English as a *lingua franca* (ELF), which focuses on the use of English by speakers who do not share a language background, has gained momentum in recent years (e.g. Seidlhofer 2001 and Seidlhofer 2010; Mauranen 2003; Mauranen et al. 2010; House 2013) and acknowledges the wide variety of speech situations in which ELF is used. For example, English can be the means of communication in very formal settings, such as business negotiations or academic lectures. In these speech situations, the focus is on the exchange of information and the language will have an informational

***Corresponding author: Huib Kouwenhoven,** Radboud University, Centre for Language Studies, Nijmegen, The Netherlands, E-mail: huib.kouwenhoven@gmail.com
Mirjam Ernestus, Radboud University, Centre for Language Studies & Max Planck Institute for Psycholinguistics, E-mail: m.ernestus@let.ru.nl
Margot van Mulken, Radboud University, Centre for Language Studies,
E-mail: m.v.mulken@let.ru.nl

character (e.g. Biber et al. 1998). In addition, ELF is used in informal settings, such as get-togethers of international exchange students. In these settings, the focus is on involved, interactive language (e.g. Biber et al. 1998). Importantly, Firth (2009: 164) notes that in the international business encounters he studied, a pattern of 'small talk' preceding 'work talk' is observable, suggesting that non-native (L2) speakers may engage in both an informal, involved and a formal, informational speech situation within one single encounter.

This raises the question whether non-native users of English adapt their language to the formality of the speech situation, in particular when they only communicate with other non-native users of English and no native speakers are present who could set a certain norm. We contribute to answering this question by investigating whether Spanish speakers of English, who are involved in an ELF communicative setting with Dutch speakers of English, show register variation. In order to answer this question, we have developed a new corpus of non-native speech, which will also be presented in this paper.

Ample investigations of native (L1) speakers have shed light on the variability of language use according to the speech situation. We know from these studies that L1 speakers adapt their language use to the situational context by varying word choice, pronunciation and syntactic structures, for example (e.g. Biber 1988; Biber and Conrad 2009; Ernestus et al. 2015; Lee 2001; Van Herk 2012). This adaptation to the speech situation has been studied in different languages. For instance, as described by Biber and colleagues (Biber 1988; Biber et al. 1998 and Biber et al. 2006), native speakers of both English and Spanish use first and second person pronouns, causative subordination and present tense verbs more often in spontaneous conversations than in formal interviews and written language. Informational discourse, including academic writing and to a lesser extent formal interviews, is characterized by a high word type/word token ratio, longer words, more (premodifying) attributive adjectives and more nouns (Biber 1988; Biber et al. 1998 and Biber et al. 2006).

Analyses of register variation by speakers of an L2 are very few, but difficulties with situational variation may be expected. Thompson and Brown (2012) put forward that register variation may be acquired late, only after more basic language skills, such as grammar and oral expression. Moreover, even if L2 users do have the knowledge about variation, they can still encounter difficulties remembering and applying all characteristics of a given register simultaneously (Dewaele and Wourm 2002). For example, when focusing on producing grammatically correct language, an L2 speaker may lose track of the appropriate pronunciation forms given the speech situation. These difficulties may be due to the gap between the acquisition of linguistic forms and their socially appropriate use. Kecskes and Papp (2000) state that children simultaneously acquire

knowledge about linguistic forms and their socially appropriate use in their L1, integrating the two types of information. In contrast, those who learn their L2 in a classroom often acquire L2 concepts with little to no information about situational context (Dewaele and Wourm 2002; Romero-Trillo 2002). As a consequence, L2 learners cannot fully develop their sociolinguistic competence (Dewaele and Wourm 2002; Romero-Trillo 2002; Geeslin and Long 2014), and they may have difficulties adapting to the speech situation.

Previous work has investigated how L2 speakers adapt their pronunciation to the situational context. These studies have shown that the influence of speech style on pronunciation is not always similar for natives and non-natives. Thompson and Brown (2012), for example, studied one very advanced Spanish speaker of English and expected a more standard pronunciation when the amount of monitoring of speech increased (following Labov 1966). They found the exact opposite: the percentage of correct articulations of the vowel /I/ deteriorated as the formality of the speech situation increased. Furthermore, Adamson and Regan (1991) compared the production of the affix *-ing* as [Iŋ] (the prestige variant in English) or [In] (the non-prestige variant) by non-native (Vietnamese and Cambodian) and native speakers of English in both monitored and unmonitored speech. The proportion of [In] was higher in unmonitored speech for male and female native speakers, and for non-native female speakers. The opposite was true for non-native male speakers, who showed a higher proportion of [In] in monitored speech. Adamson and Regan (1991) suggest that these male non-native speakers try to accommodate to a general male native English norm rather than to a situation-specific native English norm, which leads to the overuse of the casual [In] in situations where the more formal [Iŋ] is more common.

Phonology is only one aspect of language. Other linguistic variables have received less scholarly attention when it comes to L2 variation, but some studies do exist. For instance, Geeslin and Gudmestad (2008) investigated the use of indicative or subjunctive mood and of copulas in written and spoken Spanish both by native and non-native speakers. They compared written contextualized tasks (WCT; tasks that provide a context after which participants indicate their preference for some linguistic structure over another) with sociolinguistic interviews. Results showed that both native and non-native speakers of Spanish preferred the subjunctive mood over the indicative mood and *estar* over *ser* (both translated as 'to be' in English) more often in the WCT than in the interview. The researchers also found differences between the native and non-native speakers, but only for mood choice: non-natives used fewer subjunctives than natives. Dewaele (2002) studied L2 learners' use of personal pronouns in French and found that non-native speakers of French use both informal *tu* and formal *vous* but in ways that diverge from the native speaker norm. Just like the

pronunciation patterns found by Thompson and Brown (2012) and Adamson and Regan (1991), the studies by Geeslin and Gudmestad (2008) and Dewaele (2002) reveal the presence of non-native sociolinguistic competence, as reflected by the existence of systematic variation, but also differences between native and non-native variation. The consequences of this kind of deviation from the norm may be severe: it could lead to unfavorable impressions in interlocutors (Geeslin and Long 2014).

The present study extends the research on non-native register variation by investigating other, less studied, variables in two situations in which English is used by non-native speakers as *lingua franca*. First, we will investigate laughter, which previous studies have shown to be an indicator of the formality of the situation in native speech (e.g. Garcia 2013; Glenn 2010). We expect fewer occurrences of laughter in formal than in informal speech. Secondly, we will study the amount of overlapping speech, which is a measure of the high-involvement, interactive style of conversation (e.g. Tannen 2005). We expect overlapping speech to be more frequent in an informal than in a formal speech situation. Thirdly, we will analyze the number of L1 words that speakers use in their L2 English. Dewaele (2001) found that, in third language (L3) production, more L1 was used in informal than in formal speech. Following this finding, we expect more L1 words to be used in an informal than in a formal L2 English speech situation.

Then, we will test a set of 18 variables taken from the informational versus involved dimension[1] identified by Biber and colleagues (Biber 1988; Biber et al. 1998). This dimension is a scale, or continuum, on which texts can be classified based on the co-occurrence of linguistic features that share particular functions, ranging from highly informational to highly involved language, rather than a tool to indicate absolute differences between registers (Biber and Conrad 2009). The features included in our analyses are presented in more detail in Section 3.1. Based on previous research on L1 English and L1 Spanish (Biber 1988; Biber et al. 2006) we generally expect features that are characteristic of involved, interactive discourse (such as first person pronouns, second person pronouns and present tense verbs) to be used more often in informal than in formal speech. Features that are associated with informational language (such as

1 Biber (2004) also performed a factor analysis of only conversation text types. This analysis may seem more relevant for the present study since we also focus on conversational speech. However, in this more recent paper, Biber argues that the dimensions that he found to distinguish between conversation text types are strikingly similar to those he found for general spoken and written registers (Biber 1988). Since the earlier, general analysis yields more extensive descriptions of the features included in his study, we base our work on that earlier study.

nouns, long words and a high word type/word token ratio) are expected to be used less often in informal than in formal speech.

The formal and informal speech on which we base all our analyses is spontaneous speech, rather than (classroom) elicited speech. For this, we developed the Nijmegen Corpus of Spanish English (NCSE).[2] The NCSE contains conversational speech of 34 Spanish speakers of English in both a formal and an informal speech situation, in interaction with instructed Dutch confederates. We opted for Spanish and Dutch speakers of English, because Spanish belongs to a different language family than both English and Dutch. As a consequence, the issues that native speakers of Dutch and Spanish have with English in domains such as phonology and syntax are very different (see Tops et al. [2001] for Dutch and Coe [2001] for Spanish). Moreover, Spanish is not as well known in the Netherlands as French, for example. Therefore, it is less likely that Spanish and Dutch interlocutors can rely on knowledge of the other's L1.

Finally, L1 speakers of Dutch and Spanish share Western European cultural norms, and therefore are culturally determined to adapt their (language) behavior to the situational context in a similar way. To illustrate, the Official State Gazette of the Spanish government (N^{o} 178 July 2011) explicitly states that students between the ages of 6 and 12 should learn to distinguish between and to be able to produce language of different degrees of formality. Moreover, Batchelor and San José (2010) dedicate the first chapter of their reference grammar of Spanish to register variation and how register variation affects Spanish grammar. As a consequence, we may safely assume that if the Spanish speakers in the NCSE have difficulties adapting their register in English, these are linguistic rather than cultural difficulties.

The NCSE can be positioned between learner corpora and ELF corpora,[3] which both contain non-native (speech) data. Mauranen (2011) states that the main distinction between the two can be summarized by the question whether, for the speakers in the corpus, English is the object of study or a means of communication (for detailed discussions of the differences and similarities between the two types of corpora see Mauranen [2011] and Granger [2002, 2009]). ELF corpora contain naturally occurring language, authentic talk,

2 Information about how to obtain a copy of the corpus can be found at *http://www.mirjamernestus.nl/Ernestus/NCSE/index.php*.

3 An example of a learner corpus containing speech is the Louvain International Database of Spoken English Interlanguage (LINDSEI; *http://www.uclouvain.be/en-cecl-lindsei.html*). Two examples of ELF corpora are the Vienna Oxford International Corpus of English (VOICE; Seidlhofer 2010) and the Corpus of English as a Lingua Franca in Academic Settings (ELFA; Mauranen et al. 2010).

produced in real-life situations by non-native users of English. Speakers in ELF corpora, who do not share their linguistic backgrounds, use the English they master to achieve real-life goals. The NCSE shares this with ELF corpora: it includes users of L2 English whose objective was to communicate with each other, not to produce perfect English. In contrast, learner corpora comprise language from learners, who usually share their language background, and who try to acquire a certain set of (idealized, native) norms. Learner corpora are compiled following explicit design criteria and for a specific purpose, such as the study of the acquisition or the teachability of a certain linguistic feature. The NCSE was also compiled based on explicit design criteria for the purpose of collecting both formal and informal speech from the same Spanish speakers of English. However, most importantly, we tried to obtain natural language for the NCSE. We therefore tried to achieve the right balance between authenticity of the speech and ecological validity on the one hand and control over the recording quality and the degree of formality of the two speech situations on the other.

In Section 2 we will give a detailed description of the corpus creation and provide an overview of the contents of the corpus. Section 3 presents the results of our analysis of register variation based on the NCSE. In Section 4, we will discuss and interpret our results, while Section 5 provides a general discussion.

2 The Nijmegen Corpus of Spanish English

2.1 Interlocutors

As mentioned above, our study focuses on non-native speakers in situations where English is used as a *lingua franca*. For this, we included L2 speakers of English with two different L1s: native speakers of Dutch and of Spanish.

Two confederates, a 23 year old male and a 24 year old female, both undergraduate students and native speakers of Dutch, were recruited at the Radboud University. Both were selected based on their open style of communication and ability to put their interlocutors at ease. Moreover, they had ample experience with role playing in an improvisational theater group. The selection procedure of the confederates involved a short conversation in English with the first author (henceforth HK), who checked whether the candidates were proficient, but not native-like in English, in order to enhance the ecological validity of the corpus: in real-life, L2 speakers who engage in communication in English are not necessarily near-native speakers. Furthermore, the Dutch speakers of English would not be too intimidating to the Spanish speakers of English. After

the recordings of the NCSE, an experienced teacher of Cambridge ESOL/IELTS exam courses assessed the confederates' English proficiency levels at the B2/C1 level of the Common European Framework for Languages (CEFR). He did so by listening to two randomly selected excerpts of the confederates' speech. Neither of the confederates spoke Spanish. Both received payment for the two weeks of recordings.

Thirty-four Spanish university students took part in the recordings. Their ages ranged from 19 to 25 years ($M = 21.44$ years, $SD = 1.48$ years). Seventeen speakers were male, 17 were female. Most participants were near the end of their studies while two were in their first year. The majority were students of engineering, whereas five participants studied other degree subjects (law; arts; visual communications; advertising and public relations; English studies).

All Spanish participants replied to a call in which we asked volunteers to participate in a research project. This call was in Spanish, as were all other communications with the Spanish participants prior to their arrivals at the recording sessions. The call did not mention that the recordings would be in English. We proceeded in this way in order to avoid self-selection by participants based on their interest and/or proficiency in English.

The evaluator who assessed the Dutch confederates' English proficiency levels, also did so for the Spanish speakers in the NCSE: two speakers were classified at the A1 level, 10 at the A2 level, 19 at the B1 level, and three at the B2 level. An overview of the CEFR proficiency levels of the Spanish speakers in the NCSE can be found in Appendix 1.

2.2 Recording setup

The NCSE was recorded by HK in the laboratory of the *Grupo de Tecnología del Habla* at the *Escuela Técnica Superior de Ingenieros de Telecomunicación* of the *Universidad Politécnica de Madrid*. All recordings were made in a sound-attenuated room which had an approximate size of $2.80 \times 3.20 \times 3.30$ m (see Figure 1 for an overview of the setup of the recording booth during the informal setting). A large window, which overlooked the laboratory, was covered with cardboard so that HK's presence behind it would not influence the conversations. Against the wall with the window, a table was placed with on top of it several pieces of unused equipment (e.g. a PC monitor, a microphone with some cables, a camera tripod) and some cardboard boxes. Another long table was placed perpendicular to the first table and also carried some unused equipment and boxes. The interlocutors sat at this long table. The Spanish speakers were always seated at the head of the table, with the Dutch confederate sitting to their right. The

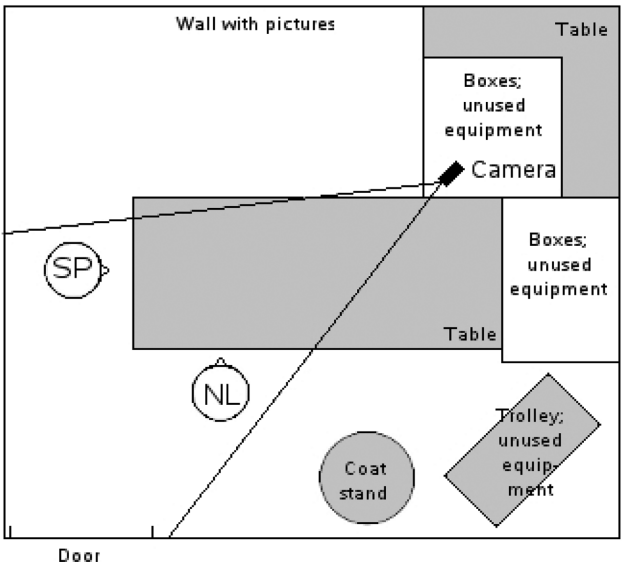

Figure 1: Setup of the recording booth in the informal setting.

walls were hung with some pictures of public figures and a map of Madrid. These could be used as conversation topics and made the room more pleasant to be in. For this reason there also was a coat rack on which the speakers could leave their coats and bags.

For the audio recordings, both speakers wore Samson QV head-mounted microphones. They were recorded in separate audio channels on an Edirol R-09 solid-state stereo recorder. The distance between the left corners of the speakers' lips and the microphones was about 3 cm. Speech signals were amplified with a stereo microphone preamplifier.

The video recordings were made by means of a Sony HDR-SR7E Handycam in HD quality (AVC HD format at 9 Mbps). During the informal part of the recordings, the camera was placed on top of a box and some cables, between the unused equipment, and with an unplugged adapter cable hanging down. The recording light of the camera was switched off. This approach effectively leaves participants unaware of the fact that they are videotaped (Torreira et al. 2010). The position of the camera was chosen so that it captured a frontal view of the Spanish participant and a side view of the Dutch confederate. For the formal part of the experiment, the camera was put on a tripod on the long table, aimed directly at the Spanish participant.

2.3 Recording procedure: informal conversation

All participants engaged in the informal part of the recordings before the formal part. As such, there was a transition from a kind of small talk in the beginning to formal communication in the end. This coincides with Firth's description of the natural development of interaction during ELF business encounters (2009).

Following Torreira et al. (2010), we tried to make the Spanish participants think that the confederate in the informal part of the recording was just another regular participant. By doing so, we created a speech situation in which the Spanish participant and the Dutch confederate were peers. Approximately ten minutes before the Spanish participant was expected to arrive, the Dutch confederate of the corresponding sex (henceforth Confederate 1) also went to the meeting point and waited for HK, as did the Spanish participant. At the agreed time, HK went out to meet the Spanish participant and Confederate 1. HK introduced himself to both and introduced them to each other. HK then asked them to wait outside while he made some final preparations. Confederate 1 was instructed to use this time to start up a conversation in order to try and break the ice.

HK started the audio and video recordings before returning to get the interlocutors. When entering the recording booth, Confederate 1 always took the same seat, leaving the chair at the head of the table for the Spanish participant. Both interlocutors were asked to put on their microphones and then HK told them that he would leave to get the task they were going to perform, and that it would be good for the project if, in the meantime, they got to know each other. HK did not explicitly mention the recordings, so that the Spanish participant would remain in doubt about whether they would start immediately or only after the speakers had received their task.

For this initial part of the informal conversation, Confederate 1 had been instructed to discretely let the Spanish participant speak most of the time. Moreover, in order to diminish the Spanish participants' potential reluctance about speaking English, Confederate 1 was instructed to make the Spanish participants feel at ease and compliment them on their English if they expressed doubts about their proficiencies.

Most conversations started with the interlocutors continuing to introduce themselves: they spoke about their education and daily lives. Quite quickly the conversations turned to other topics, such as the city of Madrid, football, travel and the crisis in Spain. This first part lasted about 25–30 minutes. When the conversation seemed to come to an end, HK returned to the recording room with a name guessing game. The interlocutors were instructed to, alternately, pick a card which had a name of a public figure (from music, cinema, politics, sports, etc.) on it. They were to describe this public figure to their interlocutor, who had

to guess the name on the card. For this part, Confederate 1 was instructed to, whenever possible, keep the conversation going about the name on the card or a related topic. This second part of the informal recordings lasted 15–20 minutes. Then, HK re-entered the recording room and invited the Spanish participant and Confederate 1 to take a short break outside the recording booth.

2.4 Recording procedure: formal interview

During the break, both the Spanish participant and Confederate 1 received written instructions, in English, about the second part of the recordings. These explained that a formal interview would be recorded as part of a graduation project for a journalism master's degree about the crisis situation in Spain and Europe. In the project's end product the interviewees' opinions would be mirrored with those of politicians and other influential people. The written instructions were aimed at putting the Spanish participants in a more formal mindset.

Once HK had changed the camera setup, placing the camera on a tripod on the table pointing it directly at the Spanish speaker, he introduced the confederate of the opposite sex (henceforth Confederate 2) to both the Spanish participant and Confederate 1. HK said that Confederate 2 was his colleague who would conduct the interviews. Confederate 2 then took the Spanish participant back into the recording booth and they both put on their microphones. HK insisted that, during the interview, the Spanish participants could freely develop their opinions and that long answers were appreciated. HK then left the recording booth.

At the beginning of the interviews, the Spanish participants formally introduced themselves, explaining their backgrounds, providing information about their families and degree programs. In the rest of the interview, most or all of the following topics were covered, but not in a fixed order: Spanish unemployment rates, government cuts on education, European pressure on Spain to cut costs, extra taxes for health care for the rich, King Juan Carlos of Spain, police attacks during student protests. As a closing act to the interview, which by that time had reached a high level of formality through the abstract nature of the topics discussed, the interviewees were asked about their expectations for their own personal life in the near and more distant future within the socioeconomic situation that they just sketched. The interview was closed after approximately 25 minutes.

The formal character of the interview was made clear in several ways. First, the camera was overtly present. Secondly, the interview was conducted by a person previously unknown to the Spanish participant. Thirdly, Confederate 2 was of the opposite sex to that of the Spanish participant. Fourthly, Confederate

2 used formal language so as to also elicit formal speech from the Spanish participant. This implied, for example, speaking clearly and not too fast, avoiding hesitations and laughter and paying attention to word choice. In addition, Confederate 2 used plural pronouns (for example *we would like to know*... rather than *I would like to know*...) in order to emphasize the idea that more people were going to watch the materials. Lastly, Confederate 2 and the Spanish participant wore formal clothing items, like a jacket, that we had asked them to bring to the recordings.

Overall, our manipulation of formality between the two parts of the recordings involved four of Biber's (1988; his terminology in italics) eight main components of the speech situation. First, an *audience* was added to the *communicative roles of participants*, by insisting on the fact that people other than HK and Confederate 2 would be watching the materials. Secondly, the *relation among participants* was altered: the casual peer to peer conversation in the informal recording was changed into an interview in which Confederate 2 had the lead. Thirdly, the *setting* was changed by adding a *superordinate activity type*: in contrast to the informal conversation, which was not linked to any other speech event, the formal interview was presented as part of a bigger entity, namely a graduation project. Lastly, the *topic* was free in the informal conversations but restricted and limited to serious issues in the formal interview.

2.5 Speaker background information and informed consent

After the interview, each Spanish participant filled in a questionnaire to provide background information like age, language knowledge and education. Moreover, the questionnaire comprised evaluative items for the two parts of the recordings (e.g. about the smoothness of the communication) and for both confederates (e.g. about the interlocutor's likability and English proficiency). Participants responded to these evaluative items on seven point Likert scales.

Once the Spanish participants had completed the questionnaires, HK provided details about the objectives of the recordings. He also made clear that the camera had been rolling during both parts of the recordings and that both confederates had been instructed beforehand. When the Spanish participants indicated their understanding of the procedure, they were asked whether they had any objections against this procedure and/or the use of the materials recorded. At this point, they were free to withdraw their personal recorded material, but none did so. All participants signed consent forms stating that the recorded materials could be used for academic purposes. They received financial rewards for their participation.

2.6 Orthographic transcription

The corpus was orthographically transcribed in PRAAT (Boersma and Weenink 2012). A transcription manual was developed specifically for the NCSE, based on previous work by MacWhinney (2000) and Torreira et al. (2010). The speech of every recording was transcribed in a PRAAT TextGrid file with three tiers: one for the Spanish speaker, one for the Dutch speaker and one for background information, for example to indicate background noise or to denote moments when HK gave instructions (see Figure 2 for an example).

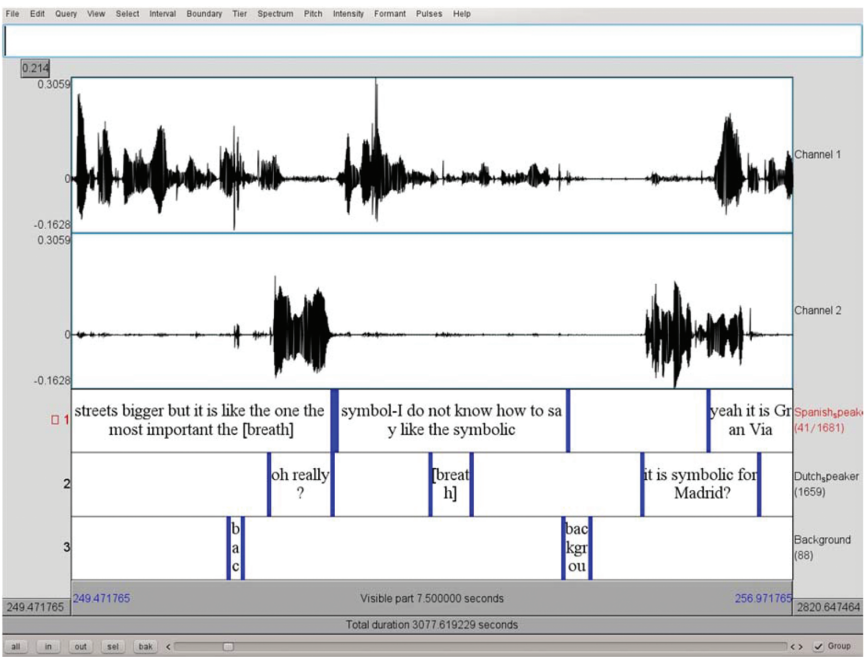

Figure 2: Screenshot of a transcription in PRAAT.

The speech was segmented into chunks with a mean length of approximately two seconds, containing on average 4.2 words. Because the chunks are that short, the orthographic transcription is well aligned with the speech signal, which facilitates finding a lexical item in this acoustic signal. Moreover, the short chunks of orthographically transcribed speech, in combination with a good pronunciation dictionary and phone models, can be used to automatically generate phonetic transcriptions.

The transcriptions were made in standard American English spelling. Contractions, such as *don't*, were written in full (*do not*). Some particular speech tokens could not be transcribed in standard American English, for example Spanish or Dutch words or truncated words. These words were annotated with special symbols, an overview of which can be found in Table 1. Frequently recurring noises, such as breaths and laughter, were transcribed between square brackets, for example *[breath]* and *[laughter]*. If words were uttered during laughter, the start and the end of the laughter were indicated, as in *[start laughter] ok it is easy [end laughter]* (for two examples of what these transcriptions look like, see Appendix 2).

Table 1: Transcription symbols used in the NCSE.

Event type	Symbol	Example
Spanish words	*	*si
Dutch words	**	**ja
Other language	***	***Deutschland
Pronunciation error	^	^Barsil (for Brazil)
Words for sounds	#	#tu #tu #tu
Spanish word made English	*^	*^aficionate
Truncated words	\-	if you go out eh abou\- eh of the s\- the school
Unintelligible speech	xxx	and it is xxx you eh

2.7 Corpus contents: lab speech or authentic talk?

Table 2 gives an overview of the duration of the recorded speech and the total number of words in the NCSE. It shows that the Spanish participants talked more

Table 2: Contents of the NCSE: duration of speech, and numbers of word types and word tokens. The type and token counts do not include truncated words.

Total duration of speech	38 h 29 min
Duration of speech in informal setting	25 h 13 min
Dutch confederates	10 h 8 min
Spanish participants	15 h 5 min
Duration of speech in formal setting	13 h 16 min
Dutch confederates	3 h 39 min
Spanish participants	9 h 37 min
Total number of word tokens (Spanish speakers only)	229,415
Total number of word types (Spanish speakers only)	6,411

than the Dutch confederates. Moreover, it reveals that the NCSE contains about two times more informal than formal speech.

We have checked participants' perception of the *naturalness* of the speech in the recordings, which we define here as a measure of how authentic or natural the speakers believed the talk to be, despite the laboratory setting. Our notion naturalness incorporates smoothness, spontaneousness and pleasantness of the communication, among others, measured by five items in our questionnaires ('The conversation/interview went well', 'The conversation/interview went smoothly', 'The conversation/interview was spontaneous', 'The conversation/interview was easy', 'The conversation/interview was pleasant'; these are translations of the Spanish items). The internal consistency of these five items was excellent for the informal ($\alpha = 0.92$) and good for the formal ($\alpha = 0.83$) setting. We therefore averaged over these five variables to create a single variable expressing naturalness.

Importantly, the talk in both the informal and the formal speech situation was reported to be natural, as shown by the mean evaluations, which were on the higher side of the seven point Likert scale ($M_{\text{formal}} = 5.31$, $\text{SD} = 1.13$; $M_{\text{informal}} = 6.19$, $\text{SD} = 1.09$). A paired t-test showed that participants' evaluations of the naturalness were significantly higher for the informal than for the formal speech situation ($t(33) = 4.84$, $p < 0.001$; Welch's approximation to the degrees of freedom was used in all t-tests in the present paper). This is as expected, given the differences between the speech situations. Overall, participants' evaluations of the naturalness, combined with the fact that the we adapted the methodology of Torreira et al. (2010), which has proven to be effective in obtaining casual speech, strengthen our belief that the speech in the NCSE can be qualified as natural.

2.8 Participants' perception of formality

We then verified whether the speakers in the NCSE were aware of the change in formality, as this was a prerequisite for all subsequent analyses. In the evaluative questionnaires, participants rated the statements 'The conversation/interview was formal'. A paired t-test showed that there was a significant effect of our formality manipulation ($t(33) = -5.03$, $p < 0.001$): the formal interviews were rated significantly more formal ($M = 5.47$, $\text{SD} = 1.42$) than the informal conversations ($M = 3.62$, $\text{SD} = 1.89$). Our manipulation has thus succeeded, which makes the NCSE a suitable collection of data to investigate whether Spanish speakers of English show register variation.

3 Register variation in the Nijmegen Corpus of Spanish English

3.1 Dependent variables and statistical analyses

In order to investigate register variation, we studied several aspects of the Spanish English speech. We compared the informal and formal parts of the NCSE on three properties of the language that previous research has put forward as indicators of speech style. We carried out these comparisons by means of linear mixed effect models with speaker as a random factor and formality as the main fixed predictor. We also checked whether the effect of formality varied per speaker (i.e. whether the random slope for formality by speaker was significant). Since we analyzed three dependent variables, we applied a Bonferroni correction and set our α-level at .017.

In some models, we added other control variables, which we will indicate below. Proficiency level was a control variable that we intended to include in all our models, but we could not do so. The proficiency data available are the CEFR scores of the speakers in the NCSE. These scores are categories, rather than values on a continuous scale, and the speakers are divided very unequally over the proficiency scores observed (see Appendix 1), which prevented us from including proficiency in our models.

First, we looked at the amount of laughter. We analyzed a relative measure for laughter expressing the mean number of laughs per 100 seconds (La/100s).

Secondly, we analyzed the amount of overlapping speech produced by each Spanish speaker. We only considered instances where the Spanish speaker interrupted the Dutch confederate, not the other way around. We calculated the amount of overlap by adding up the durations of the stretches of speech produced by the Spanish speaker while the Dutch confederate was still speaking. In this analysis, we controlled for the total duration of the speech produced within one recording by the Spanish speaker, since we expected that the more speech he or she produced, the greater the amount of overlap would be. Because this total duration of speech was significantly higher in the informal conversations ($M = 1{,}604.20$ s, $SD = 334.70$ s) than in the formal interviews ($M = 1{,}019.49$ s, $SD = 207.24$ s), we orthogonalized total duration and formality: not the raw total duration was included as a co-variate in the analysis, but the residuals of a linear regression model that predicted total duration as a function of formality.

Thirdly, we analyzed the total number of Spanish words in each recording. Since these numbers were not normally distributed, we reduced the skewness in

the data by taking the log of the number of Spanish words, which was then included as the dependent variable. In this analysis, we controlled for the total number of words in each recording, since we expected more Spanish words if the total number of words was higher. Given that there were significantly more words in the informal ($M = 4{,}069.62$, $SD = 1{,}098.58$) than in the formal ($M = 2{,}677.59$, $SD = 866.09$) recordings, we orthogonalized the variables formality and total number of words: instead of including the raw number of total words in the analysis, we included the residuals of a linear regression model that predicted the total number of words as a function of formality.

Next, we examined all linguistic features that Biber and colleagues (Biber 1988; Biber et al. 1998) identified on the involved versus informational dimension and that we were able to test on the basis of the NCSE (i.e. that did not require information about punctuation or contracted forms, for example). These 18 features are listed in Table 3.

Table 3: The 18 variables from Biber and colleagues (Biber 1988; Biber et al. 1998) that were included in the present study.

Features characteristic of involved language	
Second person pronouns	Private verbs
'Be' as main verb	Demonstrative pronouns
The pronoun 'it'	First person pronouns
Possibility modals	Indefinite pronouns
Emphatics/Amplifiers	*Wh*-clauses
Verbs in the present tense	*Wh*-questions
Causative subordination	
Features characteristic of informational language	
Attributive adjectives	
Nouns	
Prepositional phrases	
Long words	
High word type/word token ratio	

We investigated whether, as predicted, the formal interviews contained more nouns, prepositional phrases and attributive adjectives than the informal conversations and whether the words were longer and the word type/word token ratio was higher in the formal interviews than in the informal conversations. These features all indicate 'a high informational focus and a careful integration of information in a text' (Biber 1988: 104).

Furthermore, we examined whether the informal conversations show higher frequencies of the thirteen involved features listed in Table 3 than the formal

interviews. We will now shortly explain why, according to Biber (1988), these features are characteristic of involved language, printing their names in italics. The *pronoun 'it'*, *indefinite pronouns* (e.g. *anybody*, *everyone*, *somebody*) and *demonstrative pronouns* (e.g. *that, these*, *this*) substitute fuller noun phrases, hence marking a 'reduced surface form' (Biber 1988: 106). The *main verb 'be'* is characteristic of fragmented speech with predicative adjectives (e.g. *the dog is small*), as opposed to attributive adjectives (e.g. *the small dog*), which keep the information within a noun phrase. In a similar way *possibility modals* (*can*, *could*, *may*, *might*) 'mark a reduced surface form, a generalized or uncertain presentation of information, and a generally fragmented production of text' (Biber 1988: 106). Two features highlight interactive language: *second person pronouns* refer directly to the addressee, whereas *wh-questions* are primarily used when there is a specific addressee to answer them. The expression of opinions, attitudes, thoughts and emotions is also characteristic of involved language. Several features fulfill this function: *wh-clauses*, *first person pronouns*, *private verbs* (e.g. *think*, *believe*) and *causative subordination* (*because*). *Present tense verbs* refer to the immediate context of communication, hence reflecting interactiveness, and together with *private verbs* they generally mark a verbal style as opposed to a style determined by nouns. Lastly, *emphatics* (e.g. *a lot*, *really*), just as *amplifiers* (e.g. *very*, *absolutely*), are characteristic of increased feeling or involvement with the topic.

Whereas Biber (1988) presents emphatics and amplifiers as separate features, we believe that the Spanish users of English in the NCSE do not make the same distinction, but instead consider words such as *really* and *very* to have the same meaning or at least the same function. This idea is supported by an inspection of the emphatics and amplifiers produced by these speakers. Of all emphatics and amplifiers, *very* (amplifier) and *really* (emphatic) are most frequent and, importantly, the contexts in which they were used were very similar. We therefore grouped emphatics and amplifiers together in our analyses.

In his Appendix 2, Biber (1988) provides detailed explanations on how he transformed the linguistic features into rules which allowed for computer automated searches. We used these same rules to count the occurrences of these 18 linguistic features in the NCSE.

Because of the difference in total number of words between the formal and informal recordings, we analyzed standardized variables (the occurrence per 10,000 words), except for word length, for which we calculated the average word length in number of characters for each recording, and word type/word token ratio, which was calculated as the percentage of unique word types of the total number of word tokens in each recording. Since not all variables were normally distributed, we tested them with Wilcoxon signed-rank tests, which

will be reported below. If a variable was normally distributed, we also produced a linear mixed effects model, which in each case yielded comparable results. Again, we applied Bonferroni correction for multiple tests: only those differences with a $p < 0.0025$ were considered to be significant.

3.2 Laughter

We observed a fixed effect of formality on the amount of laughter ($\beta = 5.00$, $t(66) = 11.41$, $p < 0.001$): there was more laughter in the informal recordings ($M = 6.37$ La/100s, SD = 3.26 La/100s) than in the formal recordings ($M = 1.37$ La/100s, SD = 1.30 La/100s). The final LMER-model including a random slope for formality by speaker was better than a model without this random slope ($\chi^2 = 37.35$, $p < 0.001$). This reveals that the size of the effect of formality on the amount of laughter varies per speaker. The standard deviation of 2.38 La/100s for the random slope of formality by speaker reflects the variation in the size of the effect of formality for individual speakers.

3.3 Overlapping speech

As expected, we found that when the total duration of speech in a recording increased, so did the amount of overlapping speech ($\beta = 0.06$, $t(65) = 5.70$, $p < 0.001$). More importantly, our model shows that formality had an effect on the amount of overlapping speech ($\beta = 131.79$, $t(65) = 14.63$, $p < 0.001$): there was more overlapping speech in the informal recordings ($M = 166.32$ s, SD = 70.62 s) than in the formal recordings ($M = 34.53$ s, SD = 20.20 s). The final LMER-model includes a random slope for formality by speaker, because it proved to be better than a model without this random slope ($\chi^2 = 49.93$, $p < 0.001$). This shows that speakers differ in the size of the effect of formality on the amount overlapping speech. The standard deviation of 50.83 s for the random slope of formality by speaker reflects the variation in the size of the effect of formality for individual speakers.

3.4 Spanish words

In line with Dewaele's (2001) results, we found an effect of formality on the number of Spanish words ($\beta = 1.05$, $t(65) = 6.41$, $p < 0.001$). This number was higher in the informal ($M = 62.35$, SD = 185.96) than in the formal speech situation ($M = 18.88$, SD = 55.60).

The effect of the total number of words was also significant ($\beta = -0.00044$, $t(65) = -2.51$, $p = 0.014$). Interestingly, and contrary to our expectations, a higher number of total words correlated with a lower number of Spanish words. An explanation may be found in the likely correlation between the total number of words and speakers' fluencies. Since all informal and all formal recordings are approximately equally long, a lower total number of words may indicate a somewhat lower fluency in English, which may lead a Spanish speaker of English to using more Spanish words. We found support for this hypothesis through an additional analysis in which we included the number of words produced per minute, not the actual number of words produced, as a proxy of fluency: we assumed that a fluent speaker produces more words per time unit than a non-fluent speaker. We produced a linear mixed effects model predicting the number of words produced per minute as a function of the log of the number of Spanish words as a fixed factor and speaker as a random factor. The fixed effect was found to be significant ($\beta = -3.57$, $t(66) = -2.94$, $p < 0.01$). The negative β-value indicates that when the number of Spanish words increases, the number of words produced per minute decreases. So if a speaker produces more Spanish words, he or she produces fewer words per minute, which may reflect a somewhat lower fluency. Additional support for this explanation is provided by the Spearman's correlation coefficient between proficiency, as reflected by the speakers' CEFR scores, and the number of Spanish words ($r_s = -0.57$, $p < 0.001$).

3.5 Involved versus informational language characteristics

The results of the analyses of the features taken from Biber and colleagues (Biber 1988; Biber et al. 1998) involved versus informational dimension can be found in Table 4. Seven of the 18 variables differed significantly between the formal and informal speech situation in the direction we hypothesized. Four of these are informational features: as was expected, more nouns, prepositional phrases and attributive adjectives were used in the formal than in the informal speech situation and words were longer in the formal than in the informal situation. Next, as was predicted, three involved features were used more often in the informal than in the formal speech situation: second person pronouns, the pronoun 'it' and forms of 'be' as main verb.

In contrast, four of the 18 features showed significant differences in the direction opposite to what we expected. These were all involved features that were used more often in the formal than in the informal speech situation: causative subordination, possibility modals, private verbs and verbs in the present tense. We will discuss these four features, among others, in the next section.

Table 4: Results of the analyses of the 18 features taken from the involved versus informational dimension identified by Biber and colleagues (Biber 1988; Biber et al. 1998). Mean number of occurrences per 10,000 words for both speech situations (average word length in characters, word type/word token ratio in percentages) and effect sizes of the Wilcoxon signed-rank tests.

Variable	Occurrence per 10,000 words (except when indicated otherwise)		
	M_{formal}	$M_{informal}$	Effect size (r)
Significant differences, expected direction ($p < 0.001$)			
Nouns	1,170.30	935.08	.62
Prepositional phrases	793.06	629.74	.62
Attributive adjectives	187.15	149.34	.48
Word-length	3.26 characters	3.17 characters	.41
Second person pronouns	123.82	169.02	−.41
Pronoun 'it'	160.51	213.22	−.45
'Be' as main verb	152.99	240.31	−.58
Significant differences, unexpected direction ($p < 0.001$)			
Causative subordination	85.67	46.22	.54
Possibility modals	59.79	41.53	.51
Private verbs	154.25	101.13	.57
Present tense verbs	565.36	434.49	.61
Non-significant differences			
Wh-questions	4.35	8.20	–
Wh-clauses	10.86	13.08	–
First person pronouns	382.90	413.97	–
Indefinite pronouns	40.01	32.97	–
Demonstrative pronouns	32.72	31.01	–
Emphatics/Amplifiers	169.32	152.58	–
Word type/word token ratio	15.30%	15.55%	–

4 Discussion: register variation

The results above show that the Spanish speakers in the NCSE adapt their language to the speech situation. Note that for our research purposes it is more important that we found differences between the formal and informal speech situations in the NCSE than whether these differences are in the direction that we expected, mostly based on previous research with natives. The differences found show that non-natives make a distinction between formal and informal speech, whether they do so in the same way as natives is a secondary question. We will now discuss and interpret our findings.

Laughter (Garcia 2013; Glenn 2010) and overlapping speech (Tannen 2005) were both expected to occur more frequently in the informal than in the formal speech situation, and both showed such an effect, reflecting a more affective and interactive nature of the speech during the informal, peer to peer conversations. Furthermore, in line with Dewaele (2001), the number of Spanish (L1) words was higher in the informal than in the formal speech situation. This suggests that speakers' ideas about whether inserting L1 words in L2 speech is acceptable or not differ for formal and informal speech situations.

Register variation by the Spanish speakers in the NCSE is also reflected by the results of our analyses of the linguistic features taken from the involved versus informational dimension identified by Biber and colleagues (Biber 1988; Biber et al. 1998). Interestingly, especially the features that are characteristic of informational discourse present a clear picture. Four of the five informational features we tested differed significantly in the expected direction: the Spanish speakers used more nouns, more prepositional phrases and more attributive adjectives in the formal interviews and the words were longer on average. By doing so, the speakers enhanced the informational density of their language.

We found one informational feature, the word type/word token ratio, to be equal in the formal and informal speech situations, while a higher word type/word token ratio was expected in the formal interviews. Possibly, non-native speakers are hindered by their limited lexicons when trying to carefully select words that carry the intended meanings very specifically. As a consequence, they may not express nuances, but re-use the same general lexical items again and again, which leads to a low word type/word token ratio.

The analyses of the thirteen features linked to involved language show a somewhat more diffuse picture. In general, we expected these features to occur more often in the informal than in the formal speech situation. Three features met this expectation: the pronoun 'it', second person pronouns and 'be' as a main verb. Each reflects a characteristic of involved language: 'it' marks a reduced surface form by substituting fuller noun-phrases, second person pronouns allow for directly addressing the interlocutor to enhance interactiveness and 'be' as a main verb is mainly used in constructions with a predicative adjective, leading to a more fragmented way of information presentation.

Not all involved features showed a formality effect, possibly because of the positioning of the informal conversations and the formal interviews in the NCSE on the involved-informational scale: the formal interviews are more towards the informational end than the informal, peer to peer conversations, but not at the extreme end of the scale, since they still represent a spontaneous, face-to-face speech situation. Therefore, they also still show some involved characteristics. The six involved features that show no significant effect of formality are

wh-questions, *wh*-clauses, first person pronouns, indefinite pronouns, demonstrative pronouns and emphatics/amplifiers.

Contrary to our expectations based on Biber's (1988) analysis of English, four of the thirteen features linked to involved communication were used more often in the formal speech situation: private verbs, possibility modals, present tense verbs and causative subordination. We will now discuss these linguistic features in detail.

First, for private verbs and possibility modals the unexpected result may have its origin in a transfer of Spanish encoding of register variation. To recall, in English, the function of private verbs is to express opinions, attitudes, thoughts and emotions and the function of possibility modals is to express some degree of uncertainty (Biber 1988). In Spanish, the linguistic features that fulfill the same functions tend to co-occur in texts that are representative of a second dimension that Biber et al. 2006: 17) call 'spoken "irrealis" discourse'. These features include conditional tense and subjunctive mood. The text genre that has the highest score on this 'spoken "irrealis" discourse' dimension is that of political interviews, but also other spoken genres, including other types of political discourse and formal meetings, show high scores. The Spanish speakers in the NCSE possibly have attempted to produce language that they considered appropriate for a formal, politically oriented interview in which presenting opinions and some degree of uncertainty about propositions is expected. Since they could not use subjunctive mood nor conditional verb forms in English, for example, they had to rely on linguistic features that fulfill the same functions in English, such as private verbs and possibility modals. Thus, the Spanish speakers in the NCSE may have relied on their knowledge about Spanish formal discourse and used linguistic features to which Biber (1988) ascribes an involved function in English, but a particular *irrealis* function in Spanish (Biber et al. 2006). To the Spanish speakers, the functions that are fulfilled by these involved linguistic features in English are characteristic of political discourse, which makes these features appropriate during the formal interviews in the NCSE.

Secondly, our finding that causative subordination is more frequent in the formal speech situation is not surprising: in this situation the speakers more often formulated complex ideas and complex argumentation. Westin (2002) argued that causative subordination is more frequent if the key objectives of a text are argumentation, explanation and opinion defending, as is the case in the newspaper editorials she studied. This function of causative subordination is also acknowledged by Biber (1988). We therefore assume that the Spanish speakers in the NCSE rely on causative subordination to achieve the particular communicative goals of expressing complex arguments or defending views during the formal interviews.

Thirdly, according to Biber (1988), present tense verbs refer to the immediate context of communication and are therefore expected to be used more in involved than in informational speech situations. However, if the topics are all current affairs, as is the case in the formal speech situation in our study, present tense verbs are indispensable. This may explain the more frequent use of present tense verbs in the formal speech situation and, again, illustrates the Spanish speakers' way of appropriately adapting their speech to the situational context.

5 General discussion

In the present study, we investigated whether Spanish speakers of English show register variation in speech situations in which English is used as a *lingua franca*. In order to answer this question, we compiled the Nijmegen Corpus of Spanish English (NCSE), in which we manipulated the formality of the speech situation. Thirty-four Spanish speakers of English engaged in both an informal, peer to peer conversation and a formal interview with Dutch speakers of English. The Spanish speakers perceived the communication as natural in both the informal and the formal speech situations, despite the laboratory setting. Moreover, the speakers' perception of the formality of the two speech situations showed that our manipulation was successful. Consequently, the NCSE is a rich collection of formal and informal speech produced by the same Spanish users of English. The recordings are of laboratory quality and augmented by orthographic transcriptions and video recordings. These contents allow for within-speaker studies of the effect of formality of the speech situation on many (linguistic) variables and from various theoretical approaches.

Based on the NCSE, we carried out several analyses that revealed that Spanish users of English show register variation on a number of language characteristics. They laugh more, produce more overlapping speech and use more Spanish words in an informal than in a formal speech situation. Moreover, the language that they produce during an informal conversation is more interactive/involved than the language they produce during a formal interview, which is more adapted for a dense presentation of information while preserving some interactive/involved characteristics. The presence of involved linguistic features during the formal interviews can be ascribed to the fact that these are also face-to-face speech situations.

Our findings complement previous work on the effect of formality on non-native language, which had focused mostly on phonology (e.g. Adamson and Regan 1991; Thompson and Brown 2012), by investigating variation on other linguistic levels. Moreover, given the proficiency levels of the speakers in the

present study (mostly B1, with a maximum of B2, see Appendix 1), we conclude that even L2 users of English who have not (yet) reached a high proficiency level show register variation. These findings partially go against previous work on L2 register variation (Dewaele and Wourm 2002; Geeslin and Long 2014; Romero-Trillo 2002; Thompson and Brown 2012) that states that L2 sociolinguistic competence comes with higher proficiencies. Our results suggest that even at early stages of L2 acquisition some kind of sociolinguistic competence is already present.

This could have its origin in speakers' reliance on L1 sociopragmatic knowledge. Since all speakers in the NCSE have a fully developed L1 (Spanish) language system, they will also have highly developed sociolinguistic competence in their L1. Importantly, Spanish and English native speakers signal the register of their speech in similar ways: in both languages, the most important dimension of register variation opposes involved to informational language (Biber 1988; Biber et al. 2006). Moreover, the languages are similar in the linguistic features that are representative of this dimension. Consequently, Spanish speakers can rely on their intuitions based on Spanish in order to produce an appropriate speech style in English, at least when it comes to the involved-informational dimension.

It would be valuable to expand our work on register variation to ELF speakers with other mother tongues. L2 users of English with different L1s may rely on different formality conventions that exist in their L1s and apply these to their English. This may be particularly true for ELF interactions in which L1 speakers are engaged with very different cultural/linguistic backgrounds, for example speakers with a Western European L1 and speakers with an Asian L1. In these cases, besides linguistic difficulties, additional problems may arise due to cultural aspects of register variation.

Furthermore, an interesting avenue for future research would be to investigate the effect of L2 register choices on interlocutors. For instance, we have seen that the language behavior of the Spanish speakers in the NCSE generally followed predictions based on native speakers of English, but we also found that they relied more than expected on private verbs and possibility modals during the formal interviews. In Spanish, the particular functions that are fulfilled by these features are associated with formal (political) interviews, but when Spanish speakers overuse them in English formal speech, interlocutors may perceive a high level of insecurity, which could have repercussions for the image of the Spanish speakers as well (Geeslin and Long 2014).

We conclude from the present study that Spanish users of English show register variation when they speak English. They laugh more and produce more overlapping speech and Spanish words in informal than in formal speech.

Moreover the language in the formal interviews in the NCSE showed more dense information presentation than the informal, peer to peer conversations. In these latter, in contrast, the language was more focused on interactiveness than in the formal interviews. So, not only did the speakers in the Nijmegen Corpus of Spanish English perceive a difference in formality between the two recordings they participated in, but this difference was also reflected by their language behavior.

Appendix 1: Individual Spanish speakers' proficiency levels.

Male speakers	CEFR proficiency level	Female speakers	CEFR proficiency level
M1	B1–	F1	A2
M2	B1	F2	B1
M3	B1	F3	B1
M4	B1	F4	A2
M5	B1	F5	A2 +
M6	A2	F6	B1 +
M7	A2	F7	A2 +
M8	B1	F8	B1 +
M9	A2	F9	B1
M10	A2	F10	B1 +
M11	A2	F11	B1–
M12	B1	F12	B2–
M13	A2	F13	B2–
M14	B1 +	F14	B1
M15	B1 +	F15	A1
M16	A1	F16	B1–
M17	B2	F17	B1

Number of Spanish speakers by proficiency level.

CEFR proficiency level	Number of speakers
A1	2
A2	8
A2 +	2
B1–	3
B1	11
B1 +	5
B2–	2
B2	1

Appendix 2: Excerpts of formal and informal speech produced by a female Spanish speaker (SP_F2) in interaction with female Confederate 1 (Conf1; informal conversation) and male Confederate 2 (Conf2; formal interview).

Formal interview		**Informal conversation**	
SP_F2:	eh I think that the prest\- the main reason [breath] is the ^speculuc\- spe\- -/culation about the buildings [breath] people working built a lot of flat [breath] eh and it cost a lot more than the real value of this this house	SP_F2:	in Andorra
		Conf2:	*wh\- is that far?*
		SP_F2:	[breath] hm [click] near *Pirineos
		Conf2:	*[laughter] oh th\- b\- th\-*
		SP_F2:	between France and Spain
Conf1:	*hm*	*Conf2*:	*Pyrenees ok*
SP_F2:	ok? [breath] and some people [click] eh have sorry some people eh in in this moment [breath] eh I do n\- [breath] obtain a lot of money	SP_F2:	[breath]
		Conf2:	*oh yeah oh that is quite far then*
		SP_F2:	a bit
Conf1:	*hm*	*Conf2*:	*yeah I have never been skiing I do not is it do you like skiing?*
SP_F2:	ok for a work that [breath] is not eh necessary		
Conf1:	*yes*	SP_F2:	[breath] [start laughter] no no [end laughter]
SP_F2:	eh f\- eh for example		
Conf1:	*give me an example*	*Conf2*:	*no? [laughter] but did you go?*
SP_F2:	[click] [breath]	SP_F2:	no m\-
Conf1:	*give us an example*	*Conf2*:	*no*
SP_F2:	eh [click] I think that eh nurse [breath] eh it is is more important than #ts eh *^constructor	SP_F2:	but my partners hm eh hm went to this trip
		Conf2:	*your your boyfriend?*
Conf1:	*hm*	SP_F2:	partn\- no hm sorry [breath]
SP_F2:	of building ok [breath]	*Conf2*:	*eh [breath]*
Conf1:	*hm*	SP_F2:	partner
SP_F2:	and the the money which gain a nurse [breath] is e\- eh [breath] it is more small than #ts than the *^constructor ok?	*Conf2*:	*your partner*
		SP_F2:	*companeros *que *no *se *acuerdo *a *ver
		Conf2:	*is it friend?*
		SP_F2:	yes m\- my [breath] friend of class

Acknowledgements: The authors would like to thank José Manuel Pardo for putting at our disposal a sound-attenuated booth in the laboratory of the *Grupo de Tecnología del Habla* at the *Escuela Técnica Superior de Ingenieros de Telecomunicación* of the *Universidad Politécnica de Madrid*. Also, many thanks

to Juan M. Lucas Cuesta, Julian D. Echeverry and Syaheerah Lutfi for their assistance during the recordings.

Funding: This work was partly funded by an ERC starting grant (284108) to the second author.

References

Adamson, H. Douglas & Vera M. Regan. 1991. The acquisition of community speech norms by Asian immigrants learning English as a second language. *Studies in Second Language Acquisition* 13(1). 1–22.

Batchelor, R. E. & Miguel Ángel San José. 2010. *A reference grammar of Spanish*. Cambridge: Cambridge University Press.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 2004. Conversation text types: A multi-dimensional analysis. In G. Purnelle, C. Fairon, & A. Dister (eds.), *Le poids des mots: Proceedings of the 7th international conference on the statistical analysis of textual data*, 15–34. Louvain-la-Neuve: Presses Universitaires de Louvain.

Biber, Douglas & Susan Conrad. 2009. *Register, genre, and style*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad & Randi Reppen. 1998. *Corpus linguistics: Investigating language structure and use*. Cambridge: Cambridge University Press.

Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2006. Spoken and written register variation in Spanish: A multi-dimensional analysis. *Corpora* 1(1). 1–37.

Boersma, Paul & David Weenink. 2012. Praat: doing phonetics by computer [Computer program]. Version 5.3.04, http://www.praat.org/ (accessed 16 January 2012).

Boletín Oficial del Estado. Nº 178, 26 July 2011. Sec. I, 84119–84138.

Coe, Norman. 2001. Speakers of Spanish and Catalan. In Michael Swan & Bernard Smith (eds.), *Learner English: A teacher's guide to interference & other problems*, 2nd edn, 90–112. Cambridge: Cambridge University Press.

de Swaan, Abram. 2001. *Words of the world: The global language system*. Cambridge: Polity Press.

Dewaele, Jean-Marc. 2001. Activation or inhibition? The interaction of L1, L2 & L3 on the language mode continuum. In Jasone Cenoz, Britta Hufeisen & Ulrike Jessner (eds.), *Cross-linguistic influence in third language acquisition: Psycholinguistic perspectives (bilingual education and bilingualism 31)*, 69–89. Clevedon: Multilingual Matters Ltd.

Dewaele, Jean-Marc. 2002. Variation, chaos et système en interlangue française. *Acquisition et interaction en langue étrangère* 17. 143–167.

Dewaele, Jean-Marc & Nathalie Wourm. 2002. L'acquisition de la compétence sociopragmatique en langue étrangère. *Revue française de linguistique appliquée* 7(2). 139–153.

Ernestus, Mirjam, Iris Hanique & Erik Verboom. 2015. The effect of speech situation on the occurrence of reduced word pronunciation. *Journal of Phonetics* 48. 60–75.

Firth, Alan. 2009. The *lingua franca* factor. *Intercultural Pragmatics* 6(2). 147–170.

Garcia, Angela Cora. 2013. *Understanding talk in formal and informal settings*. London: Bloomsbury.

Geeslin, Kimberly L. & Aarnes Gudmestad. 2008. Comparing interview and written elicitation tasks in native and non-native data: Do speakers do what we think they do? In Joyce Bruhn de Garavito & Elena Valenzuela (eds.), *Selected proceedings of the 10th Hispanic linguistics symposium*, 64–77. Somerville, MA: Cascadilla.

Geeslin, Kimberly L. & Avizia Yim Long. 2014. *Sociolinguistics and second language acquisition: Learning to use language in context*. New York: Routledge.

Glenn, Phillip. 2010. Interviewer laughs: Shared laughter and asymmetries in employment interviews. *Journal of Pragmatics* 42(6). 1485–1498.

Granger, Sylviane. 2002. A bird's-eye view of learner corpus research. In Sylviane Granger, Joseph Hung & Stephanie Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching (language learning and language teaching 6)*, 3–33. Amsterdam: John Benjamins.

Granger, Sylviane. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. In Karin Aijmer (ed.), *Corpora & language teaching (studies in corpus linguistics 33)*, 13–32. Amsterdam: John Benjamins.

House, Juliane. 2013. Developing pragmatic competence in English as a lingua franca: Using discourse markers to express (inter)subjectivity and connectivity. *Journal of Pragmatics* 59. 57–67.

Kecskes, Istvan & Tünde Papp. 2000. *Foreign language a mother tongue*. Mahwah, NJ: Lawrence Erlbaum Associates.

Labov, William. 1966. *The social stratification of English in New York City*. Cambridge: Cambridge University Press.

Lee, David Y. W. 2001. Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology* 5(3). 37–72.

MacWhinney, Brian. 2000. *The CHILDES project: Tools for analyzing talk*, 3rd edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Mauranen, Anna. 2003. The corpus of English as lingua franca in academic settings. *TESOL Quarterly* 37(3). 513–527.

Mauranen, Anna. 2011. Learners and users – Who do we want corpus from? In Fanny Meunier, Sylvie de Cock, Gaëtanelle Gilquin & Magali Paquot (eds.), *A taste for corpora: In honour of Sylviane Granger (studies in corpus linguistics 45)*, 155–172. Amsterdam: John Benjamins.

Mauranen, Anna, Niina Hynninen & Elina Ranta. 2010. English as an academic lingua franca: The ELFA project. *English for Specific Purposes* 29. 183–190.

Romero-Trillo, Jesus. 2002. The pragmatic fossilization of discourse markers in non-native speakers of English. *Journal of Pragmatics* 34(6). 769–784.

Seidlhofer, Barbara. 2001. Closing a conceptual gap: The case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11(2). 133–158.

Seidlhofer, Barbara. 2010. Giving VOICE to English as a lingua franca. In Roberta Facchinetti, David Crystal & Barbara Seidlhofer (eds.), *From international to local English – And back again (linguistic insights: studies in language and communication 95)*, 147–164. Bern: Peter Lang.

Tannen, Deborah. 2005. *Conversational style: Analyzing talk among friends*. Oxford: Oxford University Press.

Thompson, Gregory. L. & Alan V. Brown. 2012. Interlanguage variation: The influence of monitoring and contextualization on L2 phonological production. VIAL, Vigo International Journal of Applied Linguistics 9. 107–132.

Tops, Guy A., J. Xavier Dekeyser, Betty Devriendt & Steven Geukens. 2001. Dutch speakers. In Michael Swan & Bernard Smith (eds.), *Learner English: A teacher's guide to interference & other problems*, 2nd edn, 1–20. Cambridge: Cambridge University Press.

Torreira, Francisco, Martine Adda-Decker & Mirjam Ernestus. 2010. The Nijmegen corpus of casual French. *Speech Communication* 52(3). 201–222.

van Herk, Gerard. 2012. *What is sociolinguistics?* Hoboken, NJ: Wiley-Blackwell.

Westin, Ingrid. 2002. *Language change in English newspaper editorials*. Amsterdam: Rodopi.